Executive Steering Committee For A.C.E. Policy II
(ESCAP II)
Report 20
October 11, 2001

ESCAP II: Person Duplication in Census 2000

Thomas Mule
Decennial Statistical
Studies Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The ESCAP asked us to do additional research on the 2000 Accuracy and Coverage Evaluation (A.C.E.) estimate of duplication. An inter-divisional group conducted computer matching to determine the extent of duplicate census enumerations. This analysis of duplicates is limited to the extent that there was no clerical matching and that these results are generally conservative. We were concerned that perhaps the estimate of erroneous enumerations in the A.C.E. was too low because the estimate of duplicate enumerations as measured by the A.C.E. was less than the estimate from the 1990 Post-Enumeration Survey (PES). Our matching work identified duplicate enumerations that were outside of the scope of the A.C.E. This included duplicate enumerations identified outside of the geographic search area and enumerations in housing units and group quarters outside of the A.C.E. universe.

Should ESCAP be concerned that the lower A.C.E. measure of duplicate enumerations is biasing the A.C.E. estimate of the undercount?

No, the A.C.E. measured fewer duplicate enumerations because of design differences between the A.C.E. and the PES. Our analysis found an additional 1.2 million duplicate enumerations in units that were out-of-scope for the A.C.E. but would have been in-scope for the PES. The A.C.E. estimate of duplication was different from the PES estimate because the two surveys searched for duplicate enumerations in different universes of units. Accounting for these differences produced an estimate of duplicate enumerations that was much closer to the PES estimate.

Did any patterns emerge of duplicate enumerations that were out-of-scope of A.C.E.?

Yes, there were patterns by race/ethnicity domains and age/sex categories for units in the census. There were no patterns of those units that were removed from the census.

Our matching found the following results for the race/ethnicity domains:

- For persons in housing units enumerated in the census, there were higher percentages of duplicate enumerations for both the Non-Hispanic Black and the Hispanic domains than the Non-Hispanic White or Some Other Race domain. These differences were concentrated outside the one ring of surrounding blocks of the cluster but still within the same county.

- The Non-Hispanic Black domain had a higher percentage of duplicate enumerations than the Hispanic domain between persons in housing units and persons in group quarters. The Non-Hispanic Black domain had higher amounts of duplication than the Hispanic domain between 1) persons in housing units and correctional facilities and 2) persons in housing units and college dorms.

- We saw no differences for Race/Ethnicity domain between persons enumerated in housing units in the census and those persons in housing units removed during the Census Duplicate Housing Unit process.

Our matching found the following results for the age and sex categories:

- Persons less than 30 years of age had higher percentages of duplicate enumerations than persons 30 years of age or older. We saw duplication of persons less than 30 years of age more in the area outside the one ring of surrounding blocks of the cluster but still within the same county. The duplication for persons 50 years of age or older was seen more in a different state.

- The 18-29 males and 18-29 females had higher percentages of duplicate enumerations between housing units and group quarters than the other age/sex categories. The 18-29 female duplication was predominantly in college dorms while the 18-29 males were duplicated in college dorms, correctional facilities and military group quarters.

- We saw no differences based on age or sex between persons enumerated in housing units in the census and those persons in housing units removed during the Census Duplicate Housing Unit process.

What are our overall conclusions?

In summary, the A.C.E. measure of duplicate enumerations within the search area was less than the PES estimate primarily due to design differences; therefore, it is not a concern. This report also shows that patterns of duplicate enumerations are intuitive and not unexpected. This report does not say anything about how A.C.E. treated the duplicate enumerations found in this study. This is a subject of further analysis in Feldpausch (2001b).

1. BACKGROUND

We were concerned that perhaps the estimate of erroneous enumerations in the 2000 Accuracy and Coverage Evaluation (A.C.E.) was too low because the estimate of duplicate enumerations as measured by the A.C.E. was fewer than the estimate from the 1990 Post-Enumeration Survey (PES).

1.1 Why are we concerned about erroneous enumerations and duplicate enumerations?

To estimate net coverage error, a coverage study needs to estimate the number of erroneous enumerations. One category of erroneous enumeration is persons duplicated in the census.

The PES estimated more erroneous enumerations than the A.C.E. The PES estimated that 1.6 percent of the enumerations were duplicates (Hogan 1993). This is approximately 3.97 million duplicate enumerations (Childers 2001a). The A.C.E. estimated that 0.8 percent of the enumerations were duplicates. This is approximately 2 million duplicate enumerations (Feldpausch 2001a).

1.2 Did the PES and A.C.E. provide coverage estimates for the same universes?

No. The PES estimated coverage for persons in housing units and non-institutional group quarters. Persons living in institutions, military personnel living in barracks or on ships and people living in homeless shelters were excluded in 1990 (Hogan 1993). The A.C.E. estimated coverage for persons in housing units. A.C.E. did not provide coverage of persons in group quarters (Childers 2001b).

1.3 Were all of the enumerations in Census 2000 eligible for the A.C.E.?

No. For the United States, the Census Duplicate Housing Unit operation excluded 5.9 million person records from the Census. This operation later reinstated 2.3 million of these person records in the final census count. However, none of reinstated or excluded records were part of the A.C.E. Hogan (2001) showed that the exclusion of this universe would not bias the estimate of the Dual System Estimate if the number of matches is reduced proportionately to the number of census correct enumerations. However, this could produce a lower estimate of erroneous enumerations and duplicate enumerations.

1.4 How did the Census Duplicate Housing Unit operation determine which records to reinstate?

The Census Duplicate Housing Unit operation initially identified housing units as being included in error with a relatively high likelihood based on a set of person matching and address matching rules. Their research focused on the ability of the person matching to identify duplicate housing units, rather than the duplicate person records serving as substitutions for other households.

Algorithms were established for identifying instances where a duplicate household was more likely than not to reflect a substituted enumeration, rather than a duplication of housing units (Nash 2000). These cases were among the 2.3 million person records reinstated in the census count. If these cases had been available for matching, the A.C.E. potentially may have estimated these "substituted" enumerations as duplicate enumerations if they occurred within the search area.

1.5 Were the search areas different in the PES and the A.C.E.?

Yes. The search area for duplicates in the 1990 PES was the block cluster and the ring(s) of blocks surrounding the cluster. For all non-matches or erroneous enumerations, the PES searched one or two rings of surrounding blocks depending on the type of geography. Also, the PES rematched persons in some clusters with high numbers of non-matches or erroneous enumerations. The PES extended the search area beyond two rings for some of these clusters.

The search area for the A.C.E. was primarily the block cluster. Targeted Extended Search expanded the search area for a sample of units by one ring of surrounding blocks for certain cases believed to be geocoding error.

1.6 How did we categorize the units in this analysis?

Our analysis classifies person records into the following categories based on the following types of units:

Table 1: Categories of Units in this Analysis

Category	Description
E-sample Eligible[1]	Persons enumerated in housing units that were eligible to be selected for the Enumeration sample (E sample) for the Accuracy and Coverage Evaluation.
Reinstated	Persons enumerated in housing units identified to be potential duplicates by the Census Duplicate Housing Unit process. These housing units were ineligible for the E sample and the A.C.E. matching. The Duplicate Housing Unit process examined these cases and reinstated them into the census count.
Group Quarters	Persons enumerated in group quarters
Deleted	Persons enumerated in housing units identified to be potential duplicates by the Census Duplicate Housing Unit process. These housing units were ineligible for the E sample and the A.C.E. matching. The Duplicate Housing Unit process examined these cases and did not include these in the census count.

[1] Does not include Remote Alaska

2. METHODS

This report focuses on matching census person records to determine estimates of person duplication. We implemented four steps in this analysis:

- created files for computer matching
- conducted two stages of computer matching
- created an analysis file
- produced estimates of person duplication

2.1 What files did we use for computer matching?

We created the **Source** and the **Target** files:

- The **Source file** contained the data-defined persons in E-sample eligible and reinstated housing units in the **11,303 A.C.E. sample block clusters**.

- The Target file contained the data-defined records in 1) housing units and group quarters in the census enumeration and 2) housing units deleted from the census by the Census Duplicate Housing Unit operation. The **Target File** contained **all of these records from the entire nation.**

These files contained only the necessary information for matching in order to speed processing. See Appendix A for more information on the records in the Source and Target files.

2.2 How did we do the matching?

We implemented **two stages** of computer matching. Our approach used an exact matching procedure during the first stage. This stringent approach would require records to have the same values for specified characteristics to be linked together as potential duplicates.

The second stage built on the results of the first stage. By matching persons in the first stage, we identified person duplication between two units. For the second stage, we statistically matched the persons in just these two units by using the Survey Research Division matcher. The statistical matching compares the agreement of several characteristics. We determined that two records were duplicates based on the overall agreement of those characteristics.

Because of the time constraints for this project, we were unable to clerically review the duplicate links identified by the computer matching.

2.2.1 How did we match during the first stage?

We used an exact matching approach to link duplicate records. We compared **each record on the Source file** to **every record on the Target file**.

For this exact matching, we required agreement of **all** of the following variables:

- First Name
- Last Name
- Month of Birth
- Day of Birth

To be eligible for first-stage matching, **we required each record on both files to have non-blank values for all four fields**.

While we required exact correspondence for the characteristics, we did add the following enhancements to improve the matching:

- Flip-flopped the first and last name during matching. This allowed "John Jones" to link to "Jones John".

- Removed "Jr","Sr" and "III" from the first and last name fields.

- Checked to see if the middle initial was scanned into the first or last name field. This allowed us to link "Mary L. Smith" with "Mary Smithl" or "Maryl Smith".

- Required computed age to be within one year if reported by both records.

See Appendix B for the algorithm for the first-stage matching.

2.2.2 How did we match during the second stage?

We used statistically-based matching with the Fellegi-Sunter algorithm as implemented by the Statistical Research Division at the Census Bureau**.** The strength of this approach is that it allowed us to link "Timothy" and "Tim" together. We are also able to account for data capture errors ("Steve" can be linked with "Steue"). One concern is that statistically-based matching has the potential for yielding substantially more incorrect matches than exact matching if it is applied widely. Our process of requiring an exact match during the first stage between the units minimizes this potential.

We examined the agreement of the following characteristics**:**

- First Name
- Middle Initial
- Last Name
- Month of Birth
- Day of Birth
- Computed Age

Note: We used Computed Age because a census respondent can report both their year of birth and their age on the form. The computed age accounts for the reporting of both these fields.

See Appendix C for more information on the second-stage statistical matching.

2.3 Why did we need to create the Analysis Files?

The matching files contained only the information needed to link records from the Source file to records on the Target file as duplicates. The analysis files contained each link of a Source person record to a Target person record. We appended the person, unit and block characteristics to the Source and Target person record of each link. Also, we assigned the A.C.E. sampling weights so weighted estimates of person duplication could be generated.

2.4 How did we generate estimates of person duplication?

For each link, we assign sampling weights and duplication factors. We produced estimates by summing the products of the weights and factors for various categories of interest.

For some analyses, we formed categories for:

- types of housing units (whether housing unit was counted in the census or not)
- types of group quarters
- a geographic location of the duplicate

Appendix D documents these categories.

For part of the analysis, we calculated percent duplication for two of the A.C.E. post-stratification variables: Race/Ethnicity domain and Age/Sex categories. **The denominator for these estimates was the number of data-defined persons in census housing units not including Remote Alaska.** For estimates of duplication for race/ethnicity domains or age/sex categories, we used the characteristics of the Source record. Appendix E documents the race/ethnicity domains and the denominator counts for each domain. Appendix F documents the denominator counts for each age/sex category.

For variance estimates, we used a simple jackknife methodology on the final A.C.E. cluster design. These variance estimates should be slight underestimates of the variances if they reflected the full A.C.E. cluster sampling plan.

2.4.1. Which weight did we use?

Since all of the person records in E-sample eligible or the reinstated housing units in a cluster are on the Source file, we used the cluster-level weight of the Source person.

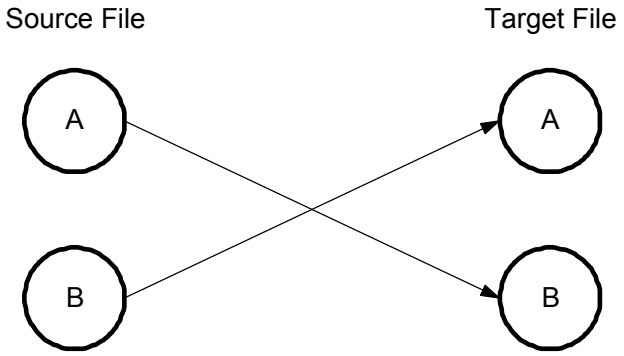
2.4.2 What factors needed to be assigned to each link?

We assigned **two factors to each link**. The first factor was an unbiased probability of duplication for the link. The second factor was a model weight which expresses the confidence in the link representing true duplication.

2.4.3 How did we assign the first factor, the unbiased probability?

The first factor was an unbiased probability of duplication for the link. A naive approach would think that each link should represent one duplicate. This would overestimate the amount of duplication when searching within the same universe (example: E-sample eligible to E-sample eligible). Here is one simple example why. In figure 1, record A is a duplicate of record B. There is only one duplicate here. Since both records are on the Source and Target files, we made two links (A to B and B to A). Thus, we need to assign each link a probability of ½ to correctly estimate one duplicate. If we assigned a probability of 1 to each link, we would have incorrectly estimated two duplicates.

Figure 1: Duplication Example



See Appendix G for more information on how we assigned the unbiased probability to each link.

2.4.4 How did we assign the second factor, the model weight?

The second factor was a model weight. This weight allows us to assign a value of confidence to the links identified in this study. This step was necessary because there was no time for a clerical review or field follow-up of the links.

We assigned a **model weight** to each link in **three parts**:

First, did our analysis identify other links between the Source unit and the Target unit?

We determined how many duplicate links were identified between the two units. The more links we identified, the more confident we were in the links.

We determined **two sets of links** where we were **confident in the links** because of the multiple links between the units. We assigned a model weight of 1 to these cases.

- All persons in the housing unit on the Source file link to the same housing unit on the Target file.

- Two or more persons in the housing unit on the Source file link to the same housing unit on the Target file within the same state.

We determined **two sets of links** that we **removed from the analysis.** These links were identified by the **second-stage matching (statistical matching)**.

- person links from housing units to group quarters. The statistical matching created too many false matches between relatives in the housing unit to other occupants of the group quarters. Example: "Margaret Brown's" sister Melanie was matched to Melanie Smith in the group quarters.

- person links between housing units in different states where the entire household was not duplicated. We were concerned about false matches when the geographic distance increased. We used state boundaries as a proxy for geographic distance.

Note: This first part assigned all of the second-stage links. The next two parts of the modeling apply to the remaining first-stage links.

Second, **do we have information to remove these cases as duplicates?**

Our processing identified the following instances where we believe the link does not represent duplication in the census.

- For links outside the cluster, the Source and Target reported different middle initials or computed ages. We allowed these links to be created in the first-stage matching to attempt to find additional links during the second-stage matching. Since we were unable to find additional links during the second stage, we removed links that had conflicting middle initials or where the computed ages differed by one year.

- Duplicate links between "Jane Doe's" and "John Doe's". These are fictitious enumerations or field imputations by the enumerator and not duplicates.

- Duplicate links with first names whose birth day is the feast day of their patron saint. We have anecdotal evidence that some people report the feast day of their patron saint as their date of birth. An example is a link between two persons named "Jose" who were born on March 19th. March 19th is the feast day of St Joseph. Appendix H lists the first names and feast day combinations which we removed from this analysis.

- Duplicate links between Nonresponse Follow-Up (NRFU) training examples. These links are fictitious enumerations and not duplicates. Appendix I lists the NRFU training example cases.

Third, for the remaining links, we have exact matches on first name, last name, month of birth and day of birth. We used a Poisson distribution approach to account for the chance that these records were linked together because of common characteristics. Our model weight compared the actual number of days with two or more births to the expected value using a Poisson distribution.

See Appendix J for more information on the modeling process.

3. LIMITATIONS

- This type of analysis has not been conducted nationally before; therefore we do not have data available for comparisons outside of the A.C.E. search areas.

- We only conducted automated matching due mostly to time constraints; there was no clerical matching or field work to resolve unknown matches. Likewise, a conservative automated matching algorithm was used to ensure that we can be confident in our identification of duplicates.

- All duplicates identified by A.C.E. were clerically identified. Clerks were able to use more characteristics and look at the scanned census forms to determine duplicates. Because of our approach, our estimate of E-sample to E-sample duplication within the cluster compared to the A.C.E. estimate will be a conservative underestimate of the duplication within this universe.

4. RESULTS

4.1 Why was the estimate of duplication in the PES different than the A.C.E.?

The A.C.E. measured fewer duplicate enumerations because of design differences between the A.C.E. and the PES.

4.1.1 What was our estimate of duplication within the cluster and the one ring of surrounding blocks?

Table 2 shows the results of our duplication analysis within the cluster and surrounding blocks for various universes.

Table 2: Person Duplication Within Cluster and Surrounding Blocks

	Within Cluster		Surrounding Block	
Universe	Estimate	Standard Error	Estimate	Standard Error
E-sample Eligible to E-sample Eligible	724,687	30,145	146,880	9,683
E-sample Eligible to Reinstated	1,049,699	41,703	24,029	6,637
Reinstated to Reinstated	15,386	4,040	1,532	542
E-sample Eligible to Group Quarter	103,168	27,820	46,736	25,595
Reinstated to Group Quarters	95	95	0	0
E-sample Eligible to Deleted	1,941,732	78,312	682,909	44,690
Reinstated to Deleted	8,767	2,796	640	334

Table 2 Highlights:

- Our estimate of duplication for E-sample Eligible to E-sample Eligible within the cluster (724,687) was 37.8 percent of the duplication for this universe identified by A.C.E.
- We identified a small number of duplicates within the cluster that were identified by our matching but missed by A.C.E. (41,046 of the 724,687). This is approximately 2 percent of the A.C.E. total estimate of duplication
- Our computer matching estimate of duplication for E-sample Eligible to Reinstated universe was very close to the clerical estimate of duplication for this universe from the Planning and Research Evaluation Division (PRED) evaluation of Reinstated persons (Raglin 2001).

4.1.2 What was the A.C.E. estimate of duplication?

Table 3 shows the A.C.E. estimate of duplication. A.C.E. searched for duplicates amongst the E-sample eligible to E-sample eligible universes.

Table 3: A.C.E. Estimate of Person Duplication

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate of E-sample Eligible to E-sample Eligible	2,014,675

Source: Feldpausch (2001a)

4.1.3 What would the estimate of duplication have been if a methodology similar to the PES had been implemented on the entire 2000 Census counts?

Table 4 shows the results of using a methodology more similar to the PES. This result is approximately 1.2 million higher than the A.C.E. estimate. These estimates extend the search area for all units to one ring of surrounding blocks. These estimates include searching for duplication to the reinstated housing units and group quarters. These housing units and the non-institutional group quarters would have been in-scope for the PES.

Table 4: Estimate of Person Duplication Using a Methodology Similar to the PES on 2000 Census Count

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate plus E-sample Eligible to Reinstate, E-sample Eligible to Group Quarters	3,238,307

4.1.4 What would the estimate of duplication have been if a methodology similar to the PES had been implemented on the 2000 Census count prior to the Duplicate Housing Unit operation?

Table 5 shows the results of using a methodology similar to the PES on the 2000 Census counts prior to the Duplicate Housing Unit operation. This result is approximately 3.8 million more duplicates than the A.C.E. estimate. This universe is not entirely comparable to the PES. Census 2000 used multiple sources of addresses when compiling the Master Address File (Nash 2000). These results show what the estimate of duplication would have been if the Duplicate Housing Unit operation was not done.

Table 5: Estimate of Person Duplication Using a Methodology Similar to the PES on a 2000 Census Count Prior to the Duplicate Housing Unit Operation

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate plus E-sample Eligible to Reinstate, E-sample Eligible to Group Quarters E-sample Eligible to Deleted	5,862,916

4.2 What was the extent of duplicate enumerations that were 1) outside of the surrounding blocks or 2) outside of the universe of A.C.E.?

4.2.1 What were the total estimates of duplication from our analysis?

Table 6 shows the estimates of duplication from our analysis for various universes. This table presents total results and results for outside the surrounding blocks. The table has two sets of estimates for the Census housing unit to Census housing unit universe. The first set includes all duplicates (Total). The second set does not include duplicate links to reinstated units. The Duplicate Housing Unit operation developed algorithms for identifying instances where a duplicate household was more likely than not to reflect a substituted enumeration, rather than a duplication of housing units (Nash 2000). Because of this, we presented both sets of estimates.

Table 6: Total Estimate of Person Duplication from Our Analysis

Universe	Estimate	Standard Error
Census Housing Units to Census Housing Units		
Total	4,625,019	77,941
Outside Surrounding Blocks	2,662,806	44,389
Not including duplicate links to reinstated units	2,960,675	47,786
Outside Surrounding Blocks	2,089,107	33,210
Census Housing Units to Group Quarters	660,189	65,119
Census Housing Units to Deleted Housing Units	2,911,016	95,665

4.2.2 What are the patterns of duplication for the Race/Ethnicity domains?

Figure K1 shows the percent duplication for the Race/Ethnicity domains for census housing units to census housing units. The figure shows similar patterns for the two sets of estimates (total and not including duplication to reinstated units). This figure shows higher percentages of duplicate enumerations for both the Non-Hispanic Black and the Hispanic domains than the Non-Hispanic White or Some Other Race domain.

Tables L1 and L2 show the percent duplication for the Race/Ethnicity domains by geography. Table L1 shows the total results using all of the duplicates identified in our analysis. Table L2 shows the results not including the duplicate links to the reinstated units. Both tables show that the Non-Hispanic Black and Hispanic duplicates outside the cluster and surrounding blocks are concentrated in the same county.

Figure K2 shows the percent duplication for the Race/Ethnicity domains for census housing units to group quarters. This figures shows a higher percentage of duplicate enumerations for the Non-Hispanic Black domain than the Hispanic domain. Table L3 shows the duplication to the type of group quarters. The Non-Hispanic Black domain had higher amounts of duplication than the Hispanic domain between 1) housing units and correctional facilities and 2) housing units and college dorms.

Figure K3 shows the percent duplication for the Race/Ethnicity domains for the census housing units to the housing units removed by the Census Duplicate Housing Unit operation. We saw no differences based on Race/Ethnicity domain between persons enumerated in housing units in the census and those persons in housing units removed during the Census Duplicate Housing Unit process.

4.2.3 What are the patterns of duplication for the Age/Sex categories?

Figure K4 shows the percent duplication for the Age/Sex categories for census housing units to census housing units. The figure shows similar patterns for the two sets of estimates (total and not including duplication to reinstated units). This figure shows higher percentages of duplicate enumerations for persons under 30 years old than for persons who are 30 years and older.

Tables L4 and L5 show the percent duplication for the Age/Sex categories by geography. Table L4 shows the total results using all of the duplicates identified in our analysis. Table L-5 shows the results not including the duplicate links to the reinstated units. Both tables show the following pattern for duplication outside of the surrounding blocks. Duplication of persons under 30 years old is concentrated more in the same county while duplication of persons 50 years and older are concentrated more in a different state.

Figure K5 shows the percent duplication for the Age/Sex categories for census housing units to group quarters. The 18-29 males and 18-29 females had higher percentages of duplicate enumerations between housing units and group quarters than the other age/sex categories.

Table L6 shows the percent duplication for the Age/Sex categories by the type of group quarters. The table shows the 18-29 female duplication was predominantly in college dorms while the 18-29 males were duplicated in college dorms, correctional facilities and military group quarters.

Figure K6 shows the percent duplication for the Age/Sex categoriess for the census housing units to the housing units removed by the Census Duplicate Housing Unit operation. We saw no differences based on Age/Sex categories between persons enumerated in housing units in the census and those persons in housing units removed during the Census Duplicate Housing Unit process.

References

Childers, D., "1990 E-Sample Documentation" Internal Census Bureau memorandum, Washington, D.C., 2001(a).

_____., "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-1, U.S. Census Bureau, Washington, D.C., 2001(b).

Feldpausch, R., "E-sample Erroneous Enumeration Analysis," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 5, U.S. Census Bureau, Washington, D.C., 2001(a).

_____., "Census Person Duplication and Corresponding A.C.E. Enumeration Status," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 6, U.S. Census Bureau, Washington, D.C., 2001(b).

Haines, D., "Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation", DSSD Census 2000 Procedures and Operations Memorandum Series Q-24, U.S. Census Bureau, Washington, D.C., 2000.

Hogan, H., "The 1990 Post-Enumeration Survey: Operations and Results" Journal of the American Statistical Association, September 1993, Volume 88 Number 423.

_____., "Accuracy and Coverage Evaluation Survey: Effect of Excluding 'Late Census Adds'," DSSD Census 2000 Procedures and Operations Memorandum Series Q-43, U.S. Census Bureau, Washington, D.C., 2001.

Nash, F., "Overview of the Duplicate Housing Unit Operations," Internal Census Bureau memorandum, Census 2000 Informational Memorandum Number 78, U.S. Census Bureau, Washington, D.C., 2000.

Raglin, D., "Effect of Excluding Reinstated Census People from the A.C.E. Person Process," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 13, U.S. Census Bureau, Washington, D.C., 2001.

Appendix A: Source and Target File

Source File

Table A1 shows the counts for the records on the Source file in this analysis. This analysis matched the E-sample eligible and the persons in reinstated housing units from this file to the Target file to estimate duplication. The person records on this file had enough characteristics to be data-defined. The E-sample eligible records were compiled from the A.C.E. PERMaRCS Census files. For the E sample cases, the clerks during A.C.E. person matching were able to update some of the names and characteristics that may have been data captured incorrectly by looking at the scanned census forms. For more information see Childers (2001b). For this analysis, we assigned the last name of the head of householder to any relative in the unit with a missing last name.

Table A1: Person Records on the Source File in This Analysis

Source	Universe
E-sample Eligible	1,820,446
Persons in the E sample	712,900
Persons sampled out of E sample	1,107,546
Reinstated Housing Units	14,561

Target File

The Target file included all of the data-defined records for the entire nation for the following units:

- E-sample Eligible
- Group Quarters
- Reinstated Housing Units
- Deleted Housing Units

Similar to the Source file, we assigned the last name of the head of householder to any relative in the unit with a missing last name.

Appendix B: First-Stage Matching

Person records on the Source and Target files were eligible for matching in the first stage if they had a first name, last name, month, and day of birth. Computed age could have been missing.

The computer algorithm

1. Read in the Source file.

2. Converted lowercase names on the ACE Source file to uppercase before setting the linked list.

3. Made a series of linked lists based on
   - day of birth,
   - month of birth,
   - initial of first name,
   - initial of last name.

   This made it easier to match.

4. Created a temporary edited first name and last name by removing non-alphabetic characters such as JR, SR, III for each Target record.

5. Read Target records and began matching within the appropriate link list from step 3.

6. Edited the last name field if the Middle initial is blank on the Source or Target file by the following:

If...	And if..	Then..
Source file's middle initial was blank	Source's last name had one more character than the last name on the Target file	set the Source's middle initial to the last character of its last name, and blanked that character out of the last name.
Target file's middle initial was blank and last name has the same number of characters on both files	Target's first name had one more character than the first name on the Source file	set the Target's middle initial to the last character of its first name, and blanked that character out of the first name.
Target file's middle initial was blank	Target's last name had one more character than the last name on the Source file	set the Target's middle initial to the last character of its last name, and blanked that character out of the last name.
Source file's middle initial was blank and last name had the same number of characters on both files	the Source's first name had one more character than the first name on the Target file	set the Source's middle initial to the last character of its first name, and blanked that character out of the first name.

7. Matched to each Source record on the appropriate linked list. Records did NOT match if
   - the edited first names were not equal, or
   - the edited last names were not equal, or
   - both ages were reported and they differed by more than 1.

8. Tallied the matches and put them on the output file. Only the unedited name went to the output file.

9. Swapped the edited first and last names if the last 2 characters of the edited last name were blank and the remaining 13 characters of the first and last names were not equal. Then repeated steps 7 and 8.

Appendix C: Second-Stage Matching

Using the results of the first stage of matching, we assigned a housing unit link identifier to all pairs of linked records. We then went back to the original Source and Target files and placed the housing unit link identifier on all person records from the housing unit regardless of their match status. Matching was then conducted using the Statistical Research Division's matcher between all persons in the linked housing units. The second stage of matching only attempted to find matches between persons in housing units where at least one match had already been identified in the first stage of matching.

We assigned matching probability weights to demographic variables before computer matching. We arrived at this set of parameters based on our past experience and knowledge of statistical matching in census operations. We matched the records based on the overall agreement of these characteristics. The second-stage matching determined two records were duplicates if the overall score was greater than the cutoff for matching.

Appendix D: Analysis Categories

D.1 Geographic Categories of the Duplicate Links

Our analysis used the following categories of geography:

Cluster and Surrounding blocks

- Within the block cluster
- With the one ring of surrounding blocks outside the cluster

Outside the cluster and one ring of surrounding blocks

- Within the same county
- Within a different county in the same state
- In a different state

D.2 Categories of Housing Units

Table D1 shows the categories of housing units in this analysis

Table D1: Categories of Housing Units

Category of Housing Units	Type of Units in Category
Census Housing Unit	• E-sample Eligible Housing Units • Reinstated
Deleted Housing Units	• Housing units removed during the Duplicate Housing Unit Operation

D.3 Categories of Group Quarters

Table D2 shows the categories of group quarters in this analysis.

Table D2: Categories of Group Quarters

Category of Group Quarters	Type of Units in Category
Correctional Institution	• Federal detection centers • Federal prisons • State prisons • Local jails • Correctional halfway houses • Military prisons • Other prisons
Nursing Homes	• Nursing home
Juvenile Institution	• Neglected/abused juvenile institutions • Emotionally distributed kids institutions • Delinquent kids institutions • Other juvenile institutions
College Dorms	• College dorms
Military	• Miliary barracks
Other	• Drug/alcohol abuse treatment • Military hospital • Civilian hospital • Hospices • Mentally ill hospital • Mentally handicapped hospital • Institution for deaf • Institution for blind • Other physically handicap • Homeless shelter • Children's shelter • Domestic violence shelter • Soup kitchen • Mobile food van • TNSOLs • Drug/alcohol group home • Mentally ill group home • Physically handicapped group home • Other group home • Agricultural worker's dorm • Other worker dorm • Job corps dorm • Staff dorms: Military hospital/prison • Religious group quarter • Hostels, YM/WCAs, etc. • Protective oversight

Appendix E: Race/Ethnicity Domains

The race/origin domain assignment generally follows the guidelines listed below, but it is essential to see Haines (2000) for the complete set of rules used to classify people into one of the seven domains. The race/origin domain assignment is hierarchical.

Domain 1 (American Indian or Alaska Native on reservations) includes:

- All people on a reservation with American Indian or Alaska Native either as their single race or as one of multiple races, regardless of their Hispanic origin.

Domain 2 (American Indian or Alaska Native off reservations) includes:

- All people in Indian Country[1] but not on a reservation with American Indian or Alaska Native either as their single race or as one of multiple races, regardless of their Hispanic origin.

- All non-Hispanic people not in Indian Country with American Indian or Alaska Native as their single race.

Domain 3 (Hispanic) includes:

- All Hispanic people in Indian Country, excluding those with American Indian or Alaska Native either as their single race or as one of multiple races.

- All Hispanic people not in Indian Country, excluding those who live in the state of Hawaii and have Native Hawaiian or Pacific Islander as a single race or as one of multiple races.

[1] Indian Country is land considered (either wholly or partially) on an American Indian reservation/trust land, Tribal Jurisdiction Statistical Area, Tribal Designated Statistical Area, or Alaska Native Village Statistical Area. For Census 2000, Tribal Jurisdiction Statistical Area has been formally renamed as Oklahoma Tribal Statistical Area.

Domain 4 (Non-Hispanic Black) includes:

- All non-Hispanic people with Black as their only race.
- All non-Hispanic people with the race combination of Black and American Indian or Alaska Native who do not live in Indian Country.
- All people with the race combination of Black and another single race group (Native Hawaiian or Pacific Islander, Asian, White, or "Some other race"), excluding those who live in the state of Hawaii and are Native Hawaiian or Pacific Islander in addition to Black.

Domain 5 (Native Hawaiian or Pacific Islander) includes:

- All non-Hispanic people with the single race Native Hawaiian or Pacific Islander.
- All non-Hispanic people with the race combination of Native Hawaiian or Pacific Islander and American Indian or Alaska Native who do not live in Indian Country.
- All non-Hispanic people with the race combination of Native Hawaiian or Pacific Islander and Asian.
- All people living in the state of Hawaii with Native Hawaiian or Pacific Islander race, regardless of their Hispanic origin and whether they identify with a single race or multiple races.

Domain 6 (Non-Hispanic Asian) includes:

- All non-Hispanic people with Asian as their single race.
- All people with the race combination of Asian and American Indian or Alaska Native who do not live in Indian Country.

Domain 7 (Non-Hispanic White or "Some other race") includes:

- All non-Hispanic people self-identifying as either White or "Some other race" as their single race, or self-identifying as both White and "Some other race."
- All non-Hispanic people with the race combination of American Indian or Alaska Native and White or "Some other race" who do not live in Indian Country.
- All non-Hispanic people with the race combinations of Asian and White or "Some other race."
- All non-Hispanic people with the race combination of Native Hawaiian or Pacific Islander and White or "Some other race," excluding those who live in the state of Hawaii.
- All non-Hispanic people with three or more races who live in Indian Country, excluding those with American Indian or Alaska Native as one of the races.
- All non-Hispanic people with three or more races and who do not live in Indian Country, excluding those who live in Hawaii and have Native Hawaiian or Pacific Islander as one of the races.

Table E1 shows the counts for the Race/Ethnicity domains. We used these counts as the denominators for the percent duplication estimates of the race/ethnicity domains. These counts are data-defined persons in housing units not including enumerations in Remote Alaska. Remote Alaska was out-of-scope for the A.C.E.

Table E1: Counts for Race/Ethnicity Domains

Race/Ethnicity Domain	Total
AI on AIR	513,147
AI off AIR	1,523,915
Hispanic	33,200,777
Non-Hispanic Black	32,330,425
Hawaiian and Pacific Islander	568,084
Non-Hispanic Asian	9,679,521
Non-Hispanic White or Some Other Race	190,004,235
Total	267,820,104

Appendix F: Age/Sex Categories

Table F1 shows the population counts for the Age/Sex categories. We used these counts as the denominators for the percent duplication estimates of the age/sex categories. These counts are data-defined persons in housing units not including enumerations in Remote Alaska. Remote Alaska was out-of-scope for the A.C.E.

Table F1: Counts for Age/Sex categories

Age/Sex Category	Total
0 - 17	69,708,968
18 - 29 Males	20,976,099
18 - 29 Females	21,024,109
30 - 49 Males	40,567,756
30 - 49 Females	42,105,085
50 + Males	33,375,084
50 + Females	40,063,003
Total	267,820,104

Appendix G: Assignment of the Unbiased Probability of Duplication

For each duplication link between the Source and Target file identified by this analysis, we need to assign an unbiased probability of duplication. We can generate a design-based estimate of duplication based on this probability. Table G1 shows the combination of duplicates we estimated in this analysis.

Table G1: Combinations of Duplicates

Combination
Duplication of E-sample Eligible to E-sample Eligible
Duplication of E-sample Eligible to GQ
Duplication of E-sample Eligible to Reinstate
Duplication of E-sample Eligible to Delete
Duplication of Reinstate to Group Quarters
Duplication of Reinstate to Reinstate
Duplication of Reinstate to Delete

Table G2 divides the records on the Source and Target files into 8 categories. The rest of this section describes how to assign probabilities based on the duplicate links between the categories.

Table G2: Categories for Assigning Unbiased Probabilities

Category	File	Description
A	Source and Target	E-sample Eligible People in A.C.E. clusters
B	Target	E-sample Eligible People not in A.C.E. clusters
C	Target	Group Quarters people in A.C.E. clusters
D	Target	Group Quarters people not in A.C.E. clusters
E	Source and Target	Reinstated People in A.C.E. clusters
F	Target	Reinstated People not in A.C.E. clusters
G	Target	Deleted People in A.C.E. clusters
H	Target	Deleted People not in A.C.E. clusters

Table G3 shows how to assign the unbiased probabilities. Each record represents a link between a Source person record and a Target person record. The Source part and the Target part of each link fall into one of the eight categories in Table G2. Since the Source records are from the A.C.E. clusters, they are in either category A (E-sample Eligible in the cluster) or category E (Reinstated person in the cluster). The table shows how to the probability is assigned based on the type of Source to Target link. Links between different universes (example: E-sample Eligible to Group Quarters) receive a probability of 1. When searching within the same universe (example: E-sample eligible to E-sample eligible), assigning a probability of 1 to each link would overestimate the amount of duplication. The table shows how to uses the number of links to other records to assign an unbiased probability. This table lists only the combinations for the estimates in our analysis.

Table G3: Assignment of Unbiased Duplication Probabilities

Source to Target Link	Duplication Probability Value
A to A	$1/(U + 1)$
A to B	$\left(\frac{1}{U + V + 1}\right)/(U + 1)$
A to C or D	1
A to E or F	1
A to G or H	1
E to C or D	1
E to E	$1/(W + 1)$
E to F	$\left(\frac{1}{W + X + 1}\right)/(W + 1)$
E to G or H	1

where U is the number of links from this Source A record to other category A records
V is the number of links from this Source A record to category B records
W is the number of links from this Source E record to other category E records
X is the number of links from this Source E record to category F records

Appendix H: First Names and Saint Feast Days Removed from Analysis

Table H1 includes the first names and birth days that were removed from this analysis. We have anecdotal evidence that some people report the feast day of their patron saint as their date of birth. We examined records with high number of links. For these records, we searched to see if the birth day was the same as the feast day of a patron saint. Most of these first names are Spanish. Links with these combinations of first name, month of birth and day of birth were assigned a model weight equal to '0'.

Table H1: First Names and Saint Feast Days Removed From Analysis

First Name	Month and Day of Birth	Saint Feast Day
Jose, Josefina	January 1	Joseph Mary Tomasi
Maria	January 1	Our Lady of Lordes
Antonio	January 17	Anthony the Abbot
Maria	February 2	Purification of Mary
Felipe	February 5	Felipe (Phillip)
Juan, Juana	March 8	John of God
Patrick, Patricia	March 17	Patrick
Jose, Josefina	March 19	Joseph
Gloria	March 25	Annunciation of the Lord
Ricardo	April 3	Richard of Chichester
Jose, Josefina	April 22	Joseph of Persia
Jorge	April 23	George
Jose, Josefina	May 1	Joseph
Cruz	May 3	Holy Cross
Isidro	May 15	Isidore of Chios
Juan, Juana	May 16	John Nepomucene
Rita	May 22	Rita of Cascia
Fernando	May 30	Ferdinand
Roberto	June 7	Robert of Newminster
Antonio	June 13	Anthony of Padua
Ismael	June 17	Ismael
Juan, Juana	June 24	John the Baptist
Alberto	June 25	St Albert of Jerusalem

Table H1: First Names and Saint Feast Days Removed From Analysis Continued...

Pedro	June 29	Peter the Apostle
Carmen	July 16	Our Lady of Mount Carmel
Jose, Josefina	July 20	Joseph Barsabas
Maria	July 22	Mary Magdalen
Cristina	July 24	Cristinia
Santiago	July 25	Santiago (James the Greater)
Ana	July 26	Ann
Clara	August 12	Clare of Assisi (Current day is August 11th)
Maria	August 15	Mary the Blessed Virgin
Luis	August 25	Luis IX
Rosa	August 30	Rose of Lima (Current day is August 23rd)
Ramon	August 31	Raymond Nonnatus
Juan, Juana	September 28	John of Cochumbuco
Miguel	September 29	Michael the Archangel
Francisco	October 4	Francis of Assisi
Eduardo	October 13	Edward the Confessor
Teresa	October 15	Teresa of Avila
Rafael	October 24	Rafael
Carlos	November 4	Carlos Borromeo
Andres	November 30	Andrew the Apostle
Concepcion, Maria	December 8	Immaculate Concepcion
Guadalupe	December 12	Our Lady of Guadalupe
Maria	December 12	Our Lady of Guadalupe
Jesus	December 24	Nativity of our Lord Jesus Christ
Jesus	December 25	Birth of Christ
Juan, Juana	December 27	John
David	December 29	David

Appendix I: Nonresponse Follow-Up Training Examples

Table I1 shows the examples used in the nonresponse follow-up training. We removed these duplicate links from our analysis because they represent fictitious enumerations.

Table I1: Fictitious Data from 2000 NRFU Enumerator Training

Householders Names	Relationship	Age	Date of Birth	Hispanic Origin	Race	Form Type	Address
Luis R. Burgos Blanca C. Burgos Chris L. Burgos Richard E. Burgos Robert L. Burgos Carlos R. Burgos Juan Burgos[2]	 wife daughter son son son brother	45 44 13 15 15 18	02/02/1955 04/10/1955 09/16/1986 06/23/1984 06/23/1984 12/22/1981 DK	Cuban for all	Black for all	Enumerator Short	4100 Herron Drive Anytown, TX 78099
Susan J. Whitman		56	10/05/1943	Puerto Rican	White	Enumerator Long	203 Forest Dr. Anytown, TX 78099
Alfred Mooney (respondent for vacant unit)							
Patrick R. Riley Ginny (No MI) Riley Rachel A. Johnson	 wife stepdaughter	49 47 13	09/15/1950 12/15/1952 06/26/1986	No	White Chinese Chinese	Enum. Short form	173 Frances Street Anycity, SC 22222
Barry L. Boswell Lorraine C. Boswell	 wife	68 66	02/16/1932 03/12/1934	No	White White	Enum. Long Form	186 Orchard Street

[2]Added in the interview as a response to a coverage question.

Appendix J: Documenting the Modeling Process

Table J1 shows the unweighted number of links used in this analysis.

Table J1: Unweighted Links in this Analysis

Source	Target	Links	Percent
E-Sample Eligible	E-Sample Eligible	116,622	71.60%
E-Sample Eligible	Group Quarters	9,618	5.90%
E-Sample Eligible	Reinstated	12,164	7.50%
E-Sample Eligible	Deleted	23,959	14.70%
Reinstated	Group Quarters	60	0.04%
Reinstated	Reinstated	322	0.20%
Reinstated	Deleted	100	0.06%
Total		162,845	

J.1 First Part of the Modeling Process

Table J2 and J3 documents the assignment of the model weight to the links based on the first part of the modeling process. The first part determined if there were other links between the Source unit and the Target unit. Table J2 documents the housing unit to housing unit links and Table J3 documents the housing unit to group quarter links. We address cases requiring further modeling in the second and third part of the process.

Table J2: Assignment of Model Weight for First Part: Housing Units to Housing Units

Analysis Category	Geography	Stage of Matching Link Made	Number of Links	Model Weight
All records in the Source unit linked to the Target unit	Within State	N/A	39,242	1
All records in the Source unit linked to the Target unit	Different State	N/A	5,648	1
Source Person for this link is in a unit where 2 or more records (but not all) in the Source unit linked to the same Target unit	Within State	N/A	15,919	1
Source Person for this link is in a unit where 2 or more records (but not all) in the Source unit linked to the same Target unit	Different State	First	5,186	Require Further Modeling
Source Person for this link is in a unit where 2 or more records (but not all) in the Source unit linked to the same Target unit	Different State	Second	5,005	0
Only one link between the Source and Target unit	N/A	Yes	82,161	Require Further Modeling
Only one link between the Source and Target unit	Same Housing Unit	No	6	1
Total			153,167	

Table J3: Housing Units to Group Quarters

Stage of Matching Link Made	Number of Links	Model Weight
First	8,141	Require Further Modeling
Second	1,537	0
Total	9,678	

J.2 Second Part of the Modeling Process

For links requiring further modeling, the question is do we have information to remove them as duplicates. The following links were given a model weight of 0. This effectively removes them from the estimates.

- For links outside the cluster, the Source and Target reported different middle initials or computed ages. We allowed these links to be created in the first-stage matching to attempt to find additional links during the second-stage matching. Since we were unable to find

additional links during the second stage, we removed these links with conflicting middle initials or computed ages which differed by one year. (51,113 links)

- Duplicate links between "Jane Doe" and "John Doe". These are fictitious enumerations or field imputations by the enumerator and not duplicates (35 links).

- Duplicate links with first names whose birth day is the feast day of their patron saint. An example is link between two persons named "Jose" who were born on March 19th. March 19th is the feast day of St Joseph (4,219 links).

- Duplicate links between Nonresponse Follow-Up (NRFU) training examples. These links are fictitious enumerations and not duplicates (347 links).

J.3 Third Part of the Modeling Process

For the remaining cases, we have exact matches on first name, last name, month of birth and day of birth. We use a Poisson distribution approach to account for the chance that these records were linked together because of common characteristics.

For each name and computed age for the link, we will compare the actual number of days with two or more births to the expected value from a Poisson distribution.

For any given name and computed age, let n denote the actual number of days with two or more census enumerations.

We then calculate the expected number of days (n^*) using a Poisson distribution by the formula below. We estimate the lambda parameter by totaling the number of births of each combination of first name, last name and computed age and dividing by 365.

If each census enumeration were unique, that is, if there were no duplicates, then

$$E(n) = n^* = 365 \sum_{t=2} \frac{\lambda^t e^{-\lambda}}{t!}$$

When $\lambda = 1/7$, for example, $n^* = 3.4$, so that roughly 3 or 4 days of multiple births are expected in any given year, in the absence of any duplicates. For $\lambda = 1$, $n^* = 96.4$.

The model weight assigned to each records is:

$$w_\lambda = \frac{n - n^*}{n}$$

The weight was negative whenever the number of observed duplicates is less than expected. We used the negative weight to reduce the bias, rather than set to 0.

Example: If there were 52 "John Smiths" with the same computed age then
$\lambda = 52/365 = 1/7$ and $n(= 3.4$.

If we counted 4 occurrences of multiple births of John Smith on the same day then the weight is:

$$\frac{4-3.4}{4} = 0.15$$

If we counted 2 occurrences of multiples births of John Smith on the same day then the weight is:

$$\frac{2-3.4}{2} = -0.70$$

(39,774 Links in the third part)

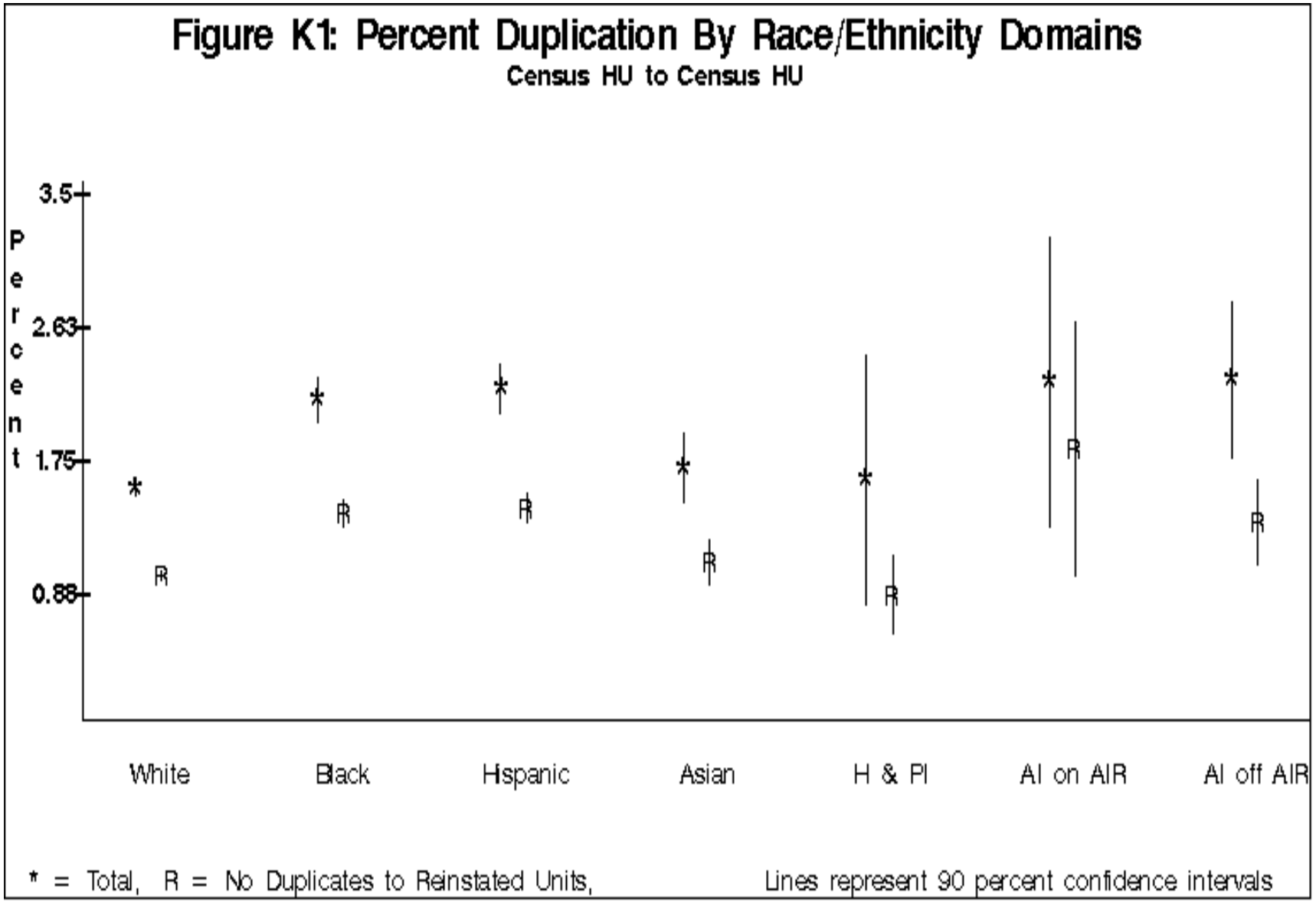


Figure K1: Percent Duplication By Race/Ethnicity Domains
Census HU to Census HU
Percent
3.5
2.63
1.75
0.88
White
Black
Hispanic
Asian
H & PI
AI on AIR
AI off AIR
* = Total, R = No Duplicates to Reinstated Units,
Lines represent 90 percent confidence intervals

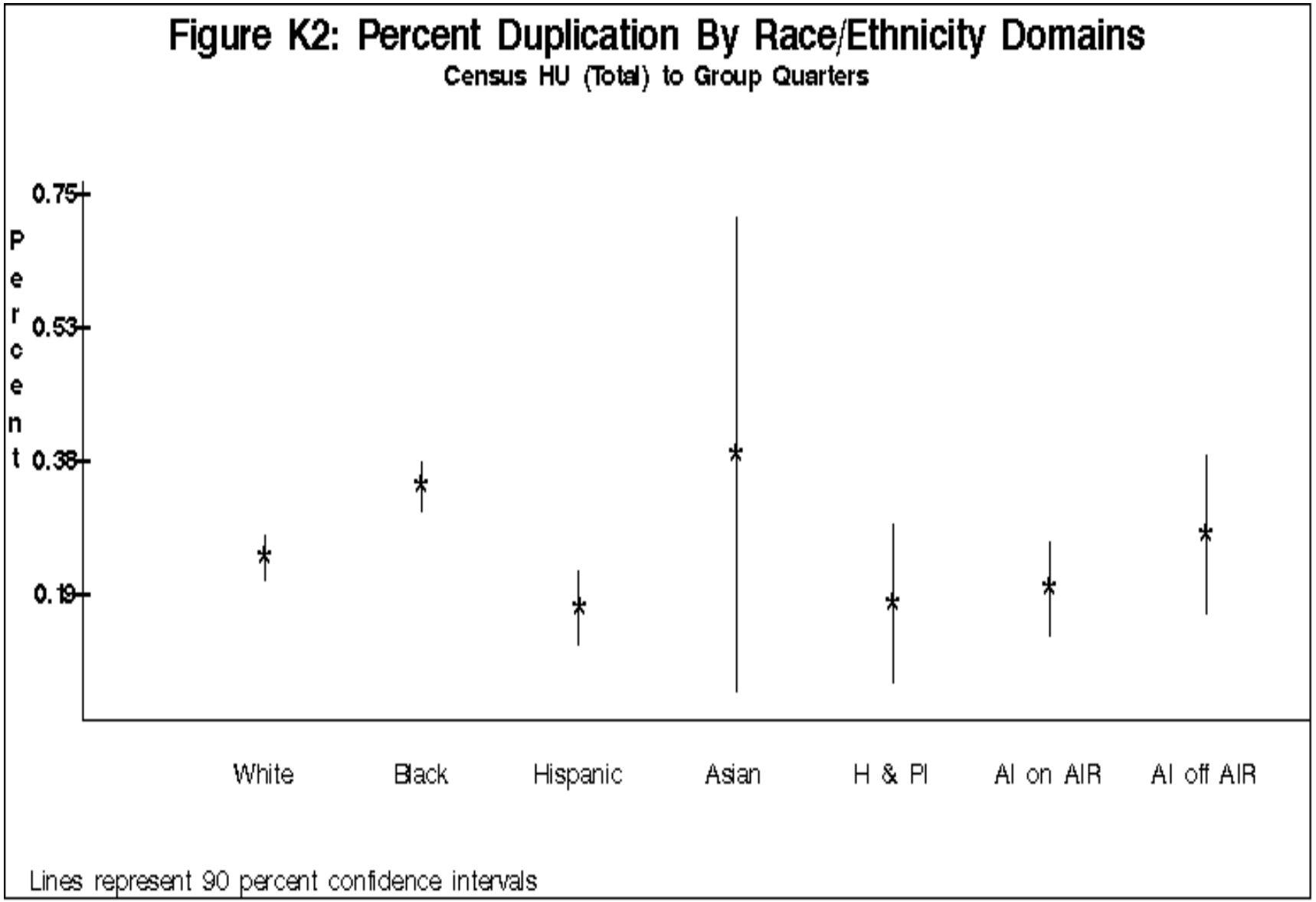


Figure K2: Percent Duplication By Race/Ethnicity Domains
Census HU (Total) to Group Quarters
Percent
0.75
0.53
0.38
0.19
White
Black
Hispanic
Asian
H & PI
AI on AIR
AI off AIR
Lines represent 90 percent confidence intervals

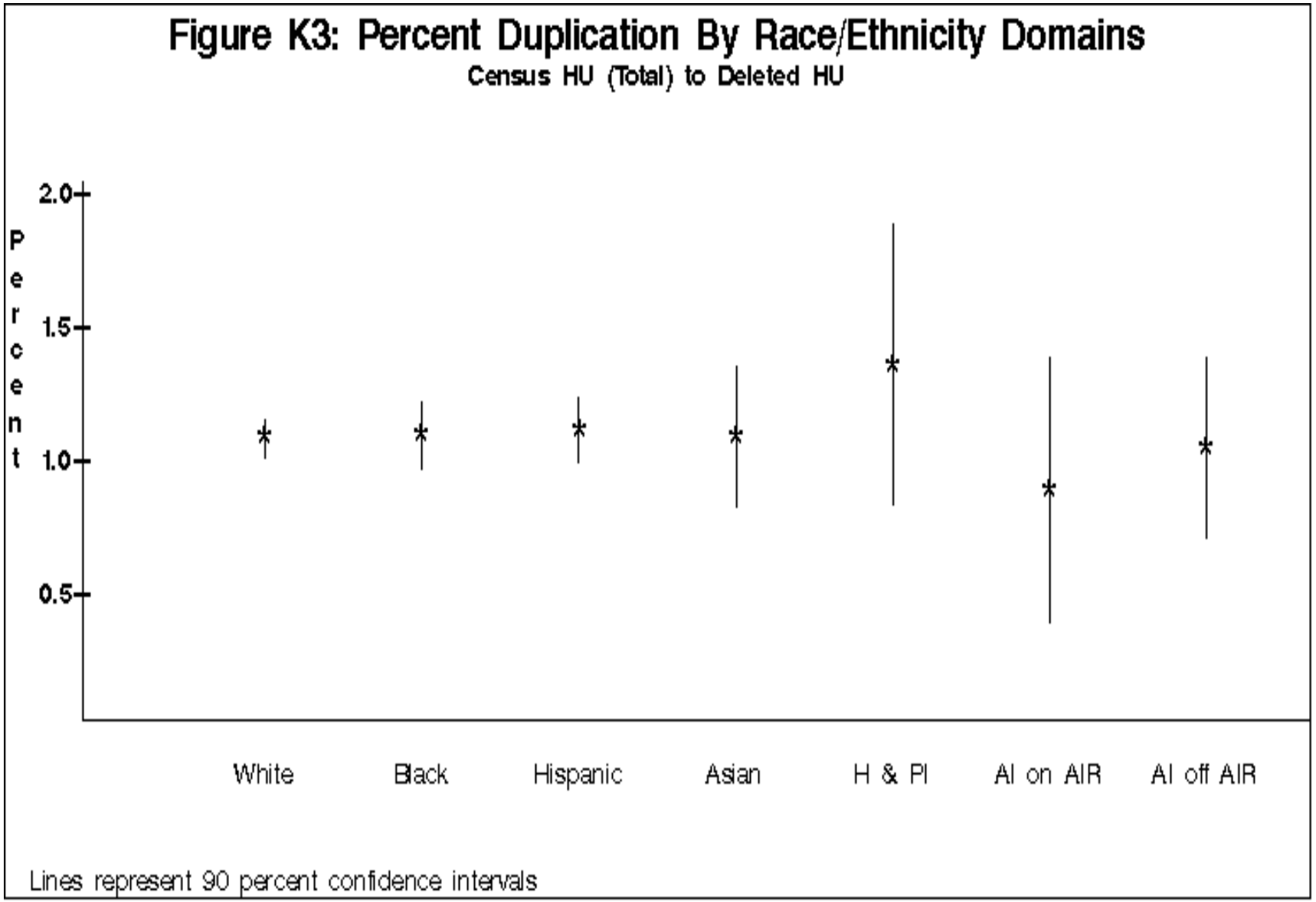


Figure K3: Percent Duplication By Race/Ethnicity Domains
Census HU (Total) to Deleted HU
Percent
2.0
1.5
1.0
0.5
White
Black
Hispanic
Asian
H & PI
AI on AIR
AI off AIR
Lines represent 90 percent confidence intervals

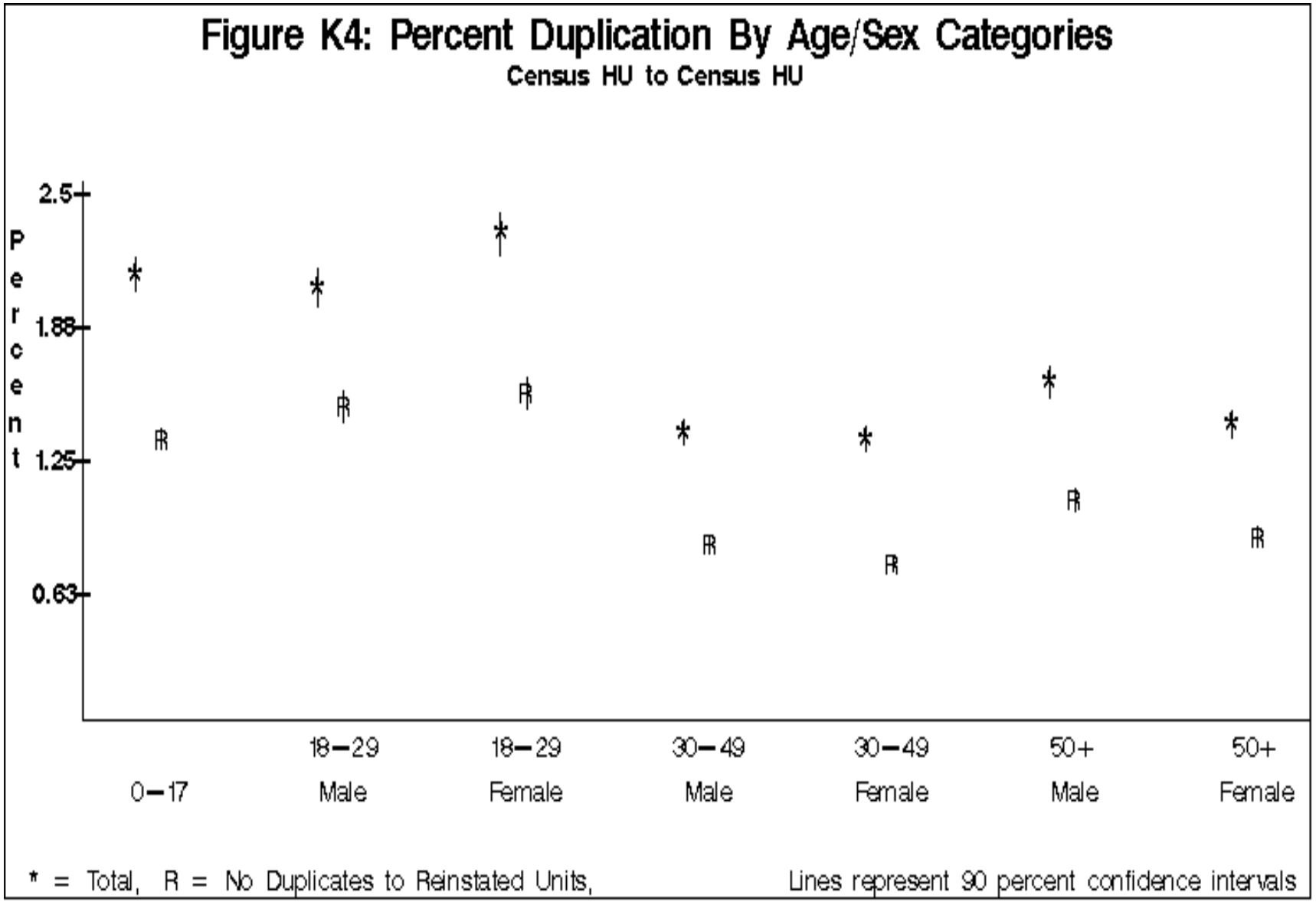


Figure K4: Percent Duplication By Age/Sex Categories
Census HU to Census HU
Percent
2.5
1.88
1.25
0.63
0–17
18–29 Male
18–29 Female
30–49 Male
30–49 Female
50+ Male
50+ Female
* = Total, R = No Duplicates to Reinstated Units,
Lines represent 90 percent confidence intervals

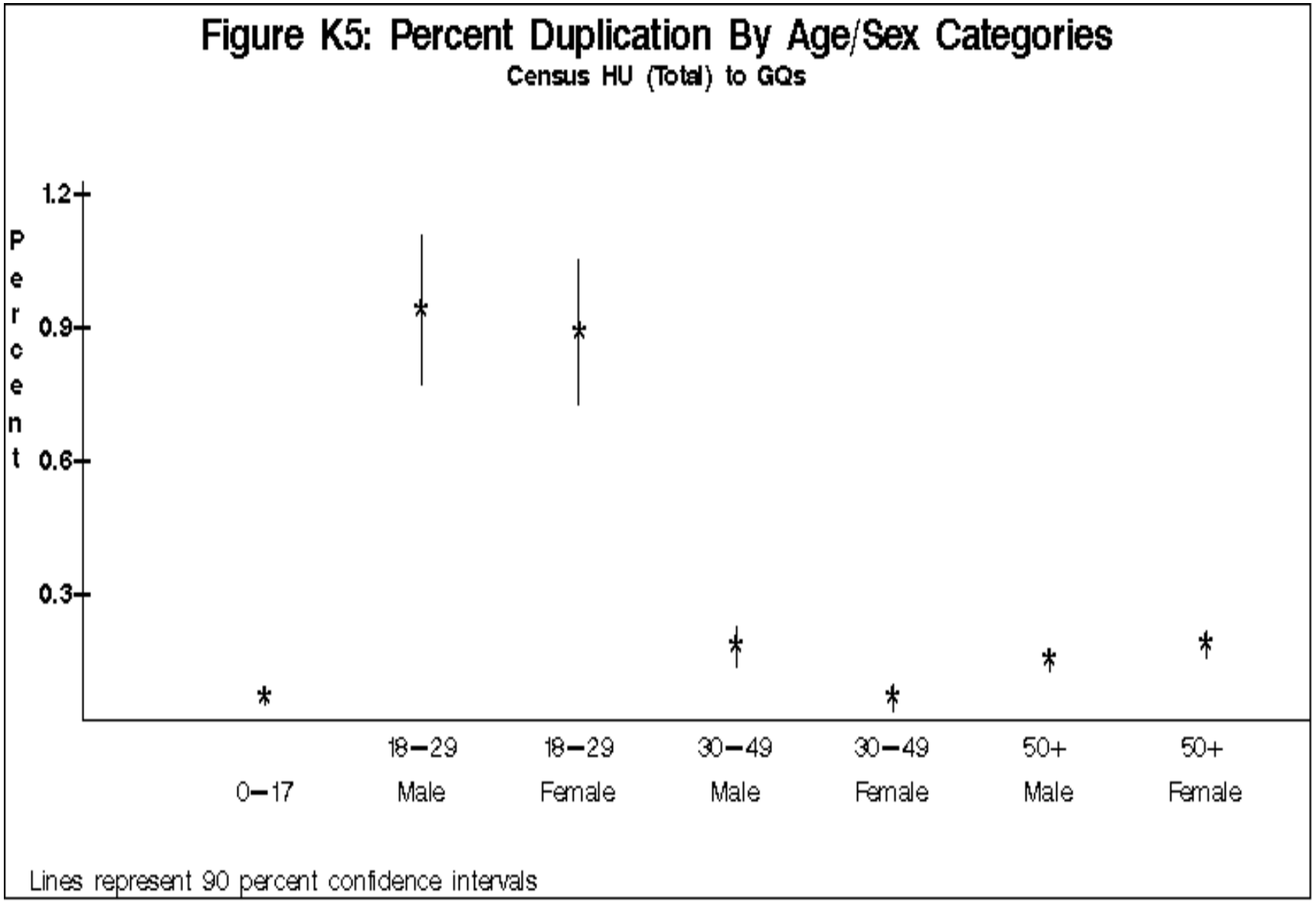


Figure K5: Percent Duplication By Age/Sex Categories
Census HU (Total) to GQs
Percent
1.2
0.9
0.6
0.3
0–17
18–29 Male
18–29 Female
30–49 Male
30–49 Female
50+ Male
50+ Female
Lines represent 90 percent confidence intervals

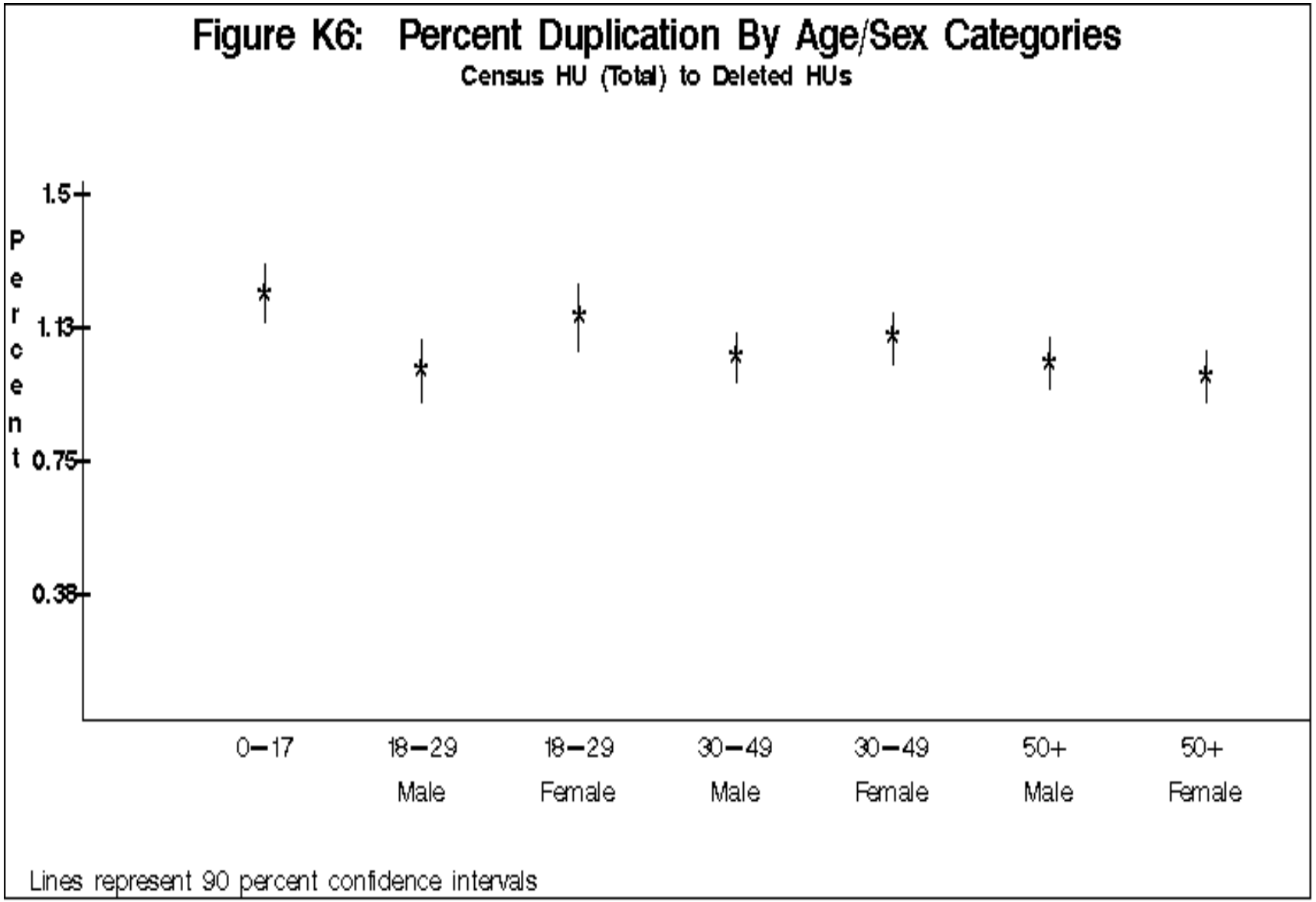


Figure K6: Percent Duplication By Age/Sex Categories
Census HU (Total) to Deleted HUs
Percent
1.5
1.13
0.75
0.38
0–17
18–29 Male
18–29 Female
30–49 Male
30–49 Female
50+ Male
50+ Female
Lines represent 90 percent confidence intervals

Table L1: Percent Duplication of Race/Ethnicity Categories by Geography
Census Housing Unit to Census Housing Unit (Total)

Race/Ethnicity Domain	Total	Within Cluster[1]	Surrounding Blocks[1]	Outside Cluster Search Area: Same County	Outside Cluster Search Area: Same State	Outside Cluster Search Area: Different State
Non-Hispanic White or Some Other Race	1.57% (0.03%)	0.56% (0.02%)	0.06% (0.01%)	0.43% (0.01%)	0.23% (0.01%)	0.29% (0.01%)
Non-Hispanic Black	2.14% (0.09%)	0.89% (0.06%)	0.09% (0.02%)	0.74% (0.04%)	0.19% (0.01%)	0.24% (0.01%)
Hispanic	2.22% (0.09%)	1.03% (0.07%)	0.06% (0.01%)	0.57% (0.03%)	0.30% (0.02%)	0.25% (0.01%)
Non-Hispanic Asian	1.70% (0.14%)	0.83% (0.12%)	0.05% (0.02%)	0.33% (0.04%)	0.26% (0.03%)	0.22% (0.02%)
Hawaiian and Pacific Islander	1.62% (0.49%)	0.41% (0.13%)	0.04% (0.02%)	0.98% (0.47%)	0.10% (0.03%)	0.09% (0.03%)
American Indian on AIR	2.26% (0.58%)	0.38% (0.11%)	0.36% (0.21%)	1.11% (0.40%)	0.23% (0.05%)	0.18% (0.03%)
American Indian off AIR	2.27% (0.31%)	0.80% (0.18%)	0.03% (0.02%)	0.69% (0.13%)	0.40% (0.13%)	0.36% (0.06%)

[1] This estimate is from our analysis and not A.C.E.

Table L2: Percent Duplication of Race/Ethnicity Domains by Geography
Census Housing Unit to Census Housing Unit (Not Including Duplicates to Reinstated Units)

Race/ Ethnicity Domain	Total	Within Cluster[1]	Surrounding Blocks[1]	Outside Cluster Search Area: Same County	Outside Cluster Search Area: Same State	Outside Cluster Search Area: Different State
Non-Hispanic White or Some Other Race	1.00% (0.02%)	0.20% (0.01%)	0.05% (0.00%)	0.26% (0.01%)	0.20% (0.01%)	0.28% (0.01%)
Non-Hispanic Black	1.40% (0.06%)	0.43% (0.04%)	0.07% (0.01%)	0.50% (0.02%)	0.16% (0.01%)	0.23% (0.01%)
Hispanic	1.44% (0.06%)	0.48% (0.04%)	0.06% (0.01%)	0.38% (0.02%)	0.28% (0.02%)	0.23% (0.01%)
Non-Hispanic Asian	1.08% (0.09%)	0.39% (0.07%)	0.05% (0.02%)	0.22% (0.02%)	0.20% (0.02%)	0.22% (0.02%)
Hawaiian and Pacific Islander	0.87% (0.15%)	0.27% (0.11%)	0.04% (0.02%)	0.38% (0.08%)	0.10% (0.03%)	0.08% (0.03%)
American Indian on AIR	1.82% (0.50%)	0.19% (0.06%)	0.35% (0.21%)	0.89% (0.30%)	0.21% (0.04%)	0.18% (0.03%)
American Indian off AIR	1.34% (0.17%)	0.31% (0.09%)	0.02% (0.01%)	0.39% (0.08%)	0.27% (0.07%)	0.36% (0.06%)

[1] This estimate is from our analysis and not A.C.E.

Table L3: Percent Duplication of Race/Ethnicity Domains by Group Quarters Type
Census Housing Unit to Group Quarters

Race/ Ethnicity Domain	Total	Correctional Institution	Nursing Home	Juvenile Institution	College Dorm	Military	Other
Non-Hispanic White or Some Other Race	0.24% (0.02%)	0.02% (0.00%)	0.04% (0.00%)	0.00% (0.00%)	0.12% (0.02%)	0.01% (0.00%)	0.04% (0.00%)
Non-Hispanic Black	0.34% (0.02%)	0.10% (0.01%)	0.01% (0.00%)	0.01% (0.00%)	0.13% (0.01%)	0.02% (0.00%)	0.06% (0.01%)
Hispanic	0.17% (0.03%)	0.03% (0.01%)	0.00% (0.00%)	0.01% (0.00%)	0.07% (0.03%)	0.01% (0.00%)	0.04% (0.01%)
Non-Hispanic Asian	0.38% (0.20%)	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)	0.34% (0.20%)	0.01% (0.00%)	0.02% (0.01%)
Hawaiian and Pacific Islander	0.17% (0.07%)	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)	0.04% (0.02%)	0.02% (0.02%)	0.11% (0.06%)
American Indian on AIR	0.19% (0.04%)	0.06% (0.02%)	0.01% (0.01%)	0.02% (0.01%)	0.03% (0.02%)	0.00% (0.00%)	0.07% (0.02%)
American Indian off AIR	0.27% (0.07%)	0.15% (0.05%)	0.01% (0.01%)	0.01% (0.01%)	0.05% (0.02%)	0.00% (0.00%)	0.05% (0.03%)

Table L4: Percent Duplication of Age/Sex Categories by Geography
Census Housing Unit to Census Housing Units (Total)

Age/Sex Category	Total	Within Cluster[1]	Surrounding Blocks[1]	Outside Cluster Search Area: Same County	Outside Cluster Search Area: Same State	Outside Cluster Search Area: Different State
0-17	2.12% (0.05%)	0.72% (0.03%)	0.08% (0.01%)	0.86% (0.02%)	0.27% (0.01%)	0.19% (0.01%)
18-29 Males	2.05% (0.05%)	0.68% (0.04%)	0.06% (0.01%)	0.66% (0.03%)	0.37% (0.02%)	0.28% (0.01%)
18-29 Females	2.30% (0.06%)	0.77% (0.04%)	0.07% (0.01%)	0.76% (0.03%)	0.41% (0.02%)	0.29% (0.02%)
30-49 Males	1.38% (0.03%)	0.65% (0.03%)	0.05% (0.01%)	0.30% (0.01%)	0.17% (0.01%)	0.21% (0.01%)
30-49 Females	1.35% (0.03%)	0.64% (0.03%)	0.05% (0.01%)	0.31% (0.01%)	0.15% (0.01%)	0.20% (0.01%)
50 + Males	1.61% (0.04%)	0.63% (0.03%)	0.06% (0.01%)	0.20% (0.01%)	0.24% (0.01%)	0.48% (0.02%)
50 + Females	1.42% (0.04%)	0.61% (0.03%)	0.05% (0.01%)	0.18% (0.01%)	0.18% (0.01%)	0.39% (0.02%)

[1] This estimate is from our analysis and not A.C.E.

Table L5: Percent Duplication of Age/Sex Categories by Geography
Census Housing Unit to Census Housing Units (Not Including Duplicates to Reinstated Units)

Age/Sex Category	Total	Within Cluster[1]	Surrounding Blocks[1]	Outside Cluster Search Area: Same County	Outside Cluster Search Area: Same State	Outside Cluster Search Area: Different State
0-17	1.34% (0.03%)	0.28% (0.02%)	0.07% (0.01%)	0.58% (0.02%)	0.23% (0.01%)	0.18% (0.01%)
18-29 Males	1.50% (0.04%)	0.35% (0.03%)	0.05% (0.01%)	0.49% (0.02%)	0.33% (0.02%)	0.28% (0.01%)
18-29 Females	1.56% (0.04%)	0.31% (0.02%)	0.06% (0.01%)	0.52% (0.02%)	0.37% (0.02%)	0.29% (0.01%)
30-49 Males	0.85% (0.02%)	0.28% (0.02%)	0.04% (0.00%)	0.18% (0.01%)	0.15% (0.01%)	0.20% (0.01%)
30-49 Females	0.77% (0.02%)	0.25% (0.01%)	0.05% (0.00%)	0.16% (0.01%)	0.12% (0.01%)	0.19% (0.01%)
50 + Males	1.06% (0.03%)	0.23% (0.02%)	0.06% (0.01%)	0.10% (0.01%)	0.20% (0.01%)	0.48% (0.02%)
50 + Females	0.89% (0.03%)	0.23% (0.01%)	0.05% (0.01%)	0.08% (0.01%)	0.15% (0.01%)	0.38% (0.02%)

[1] This estimate is from our analysis and not A.C.E.

Table L6: Percent Duplication of Age/Sex Categories by Group Quarters Type
Census Housing Unit to Group Quarters

Race/ Ethnicity Domain	Total	Correctional Institution	Nursing Home	Juvenile Institution	College Dorm	Military	Other
0 - 17	0.07% (0.01%)	0.00% (0.00%)	0.00% (0.00%)	0.02% (0.00%)	0.01%[1] (0.01%)	0.00% (0.00%)	0.03% (0.01%)
18 - 29 Males	0.93% (0.10%)	0.18% (0.02%)	0.00% (0.00%)	0.00% (0.00%)	0.61% (0.10%)	0.07% (0.01%)	0.07% (0.01%)
18 - 29 Females	0.89% (0.10%)	0.01% (0.00%)	0.00% (0.00%)		0.82% (0.10%)	0.01% (0.00%)	0.04% (0.01%)
30 - 49 Males	0.18% (0.03%)	0.09% (0.01%)	0.00% (0.00%)		0.03% (0.02%)	0.02% (0.00%)	0.05% (0.01%)
30 - 49 Females	0.07% (0.02%)	0.01% (0.00%)	0.00% (0.00%)		0.02% (0.02%)	0.00% (0.00%)	0.03% (0.01%)
50+ Males	0.15% (0.01%)	0.01% (0.00%)	0.08% (0.01%)	0.00%[1] (0.00%)	0.00% (0.00%)		0.06% (0.01%)
50+ Females	0.19% (0.02%)	0.00% (0.00%)	0.13% (0.01%)		0.00% (0.00%)		0.05% (0.01%)

[1] Estimate is not significantly different than 0.